名古屋学院大学論集　言語・文化篇　第31巻　第2号　pp. 83–96

〔レビュー・アーティクル〕

ニューラルテスト理論を用いたテスト結果のフィードバック

城　野　博　志

名古屋学院大学経済学部

要　旨

学力テストのように能力を測定する道具の性能を評価したり，能力を推定したりするための統計的な理論をテスト理論と言う。テスト理論としてこれまでに古典的テスト理論，項目応答理論，ニューラルテスト理論などが提唱されてきた。本稿は，これらテスト理論の中からニューラルテスト理論に焦点を当てて，測定と評価の観点から同理論が提供するテスト結果に対するフィードバックの先行研究を概観する。

キーワード：テスト理論，ニューラルテスト理論，フィードバック

Feedback on test results based on neural test theory

Hiroshi SHIRONO

Faculty of Economics
Nagoya Gakuin University

発行日　2020年3月31日

1. はじめに

テストデータには，測定と評価の2つの側面がある。測定の側面とは，測定道具としてのテストやテスト項目の特性に関するものである。一方，評価の側面とは，測定の結果に意味づけを行うことにより受験者の特性を明らかにすることである（中村，2002，p. 9）。テストは一度作成したらそれで終わりというものではなく，より正確な測定や適切な評価を行うためには，テスト結果を分析して，テストやテスト項目の特性を検討しながら，改良・改善を加えていかなければならない。またそのようにして測定精度が高まったテストでなければ，適切な評価はできない。

正確な測定や適切な評価を行うためには，テストが提供するデータをつぶさに検討する必要がある。テスト実施後に，テストを構成する個別の問題（項目）を分析すること（＝項目分析）を通して，テストの測定精度の改善に寄与する情報を入手することができる。また，平均値を出してテストデータを要約し，標準偏差を計算してデータのばらつきを見ることで，受験者集団の特性を知ることができる。さらに，標準得点や偏差値から集団における受験者個人の位置に関する情報も得ることができる。こうした基礎統計量は受験者集団や受験者個々に関するデータを提供してくれる。

学力テストのように能力を測定する道具の性能を評価したり，能力を推定したりするための統計的な理論をテスト理論と言う（清水等，2003）。上記の項目分析の情報や受験者特性の情報を提示してくれる代表的なテスト理論として，古典的テスト理論（classical test theory, CTT），項目応答理論（item response theory, IRT），ニューラルテスト理論（neural test theory, NTT）などがこれまでに登場している。

CTTでは，正答数に基づく得点や素点を用いて分析を行う。しかし，この理論は，次のような限界が指摘されている（大友，1996，p. 9–16）。

ⅰ．テスト問題の困難度が受験者集団に左右される
ⅱ．能力・学力の評価が受験するテストに左右される
ⅲ．テスト得点が0であっても能力がないとは言えない
ⅳ．テスト得点のどの位置でも同一の測定単位を保持することが困難
ⅴ．項目弁別力が受験者集団の特質によって相対的に決定される
ⅵ．平行テストの作成が不可能に近い
ⅶ．被験者個別の測定誤差を仮定していない
ⅷ．被験者能力に適さないテスト項目が含まれる可能性がある

こうした限界を克服するための新しいテスト理論として，IRTが言語テストの分野で注目されるようになってきた。この理論では，正答数を対数変換したロジット得点を使用することで，等間隔の目盛りを持つ比率尺度上に数値情報を示すことができる。その結果，受験者やテストの難易度に影響されることなく，受験者の能力やテスト項目の難易度を連続尺度で推定することが可

能になる（山川，2010，p. 77）。

CTTもIRTもともに学力を1点刻みの連続尺度上で評価している（荘島，2010，p. 83）。CTTは0〜100までの得点と言う連続変数を，IRTは正答数を対数変換した θ という潜在的な連続変数を仮定している。しかしながら，荘島はテストにそこまで細かい解像度があるとは考えていない。テストは学力を5〜20段階くらいにしか分けることができないと唱える。

そのような発想を起点として，荘島はNTTを開発した。NTTは，学力と言う直接観測できない心理変数を対象としている点では，IRTと同様に潜在変数モデルである（荘島，2010，p. 84）。学力を評価する尺度として，IRTは等間隔性を持った潜在的な連続変数を仮定する。それに対して，NTTは潜在ランク尺度という潜在的な順序尺度を仮定している。その最大の特徴は，1点刻みの連続した細かい値で評価するのではなく，5〜10程度の少数のランクで段階評価するところにある。NTTは，テストの結果をいくつかの段階に分ける，いわば段階評価のためのテスト理論であると言われている（木村，2016，p. 217）。荘島（2010，p. 106）はIRTに代表される連続尺度による評価とNTTに代表される順序尺度による段階評価の違いを模式的に表している（図1）。

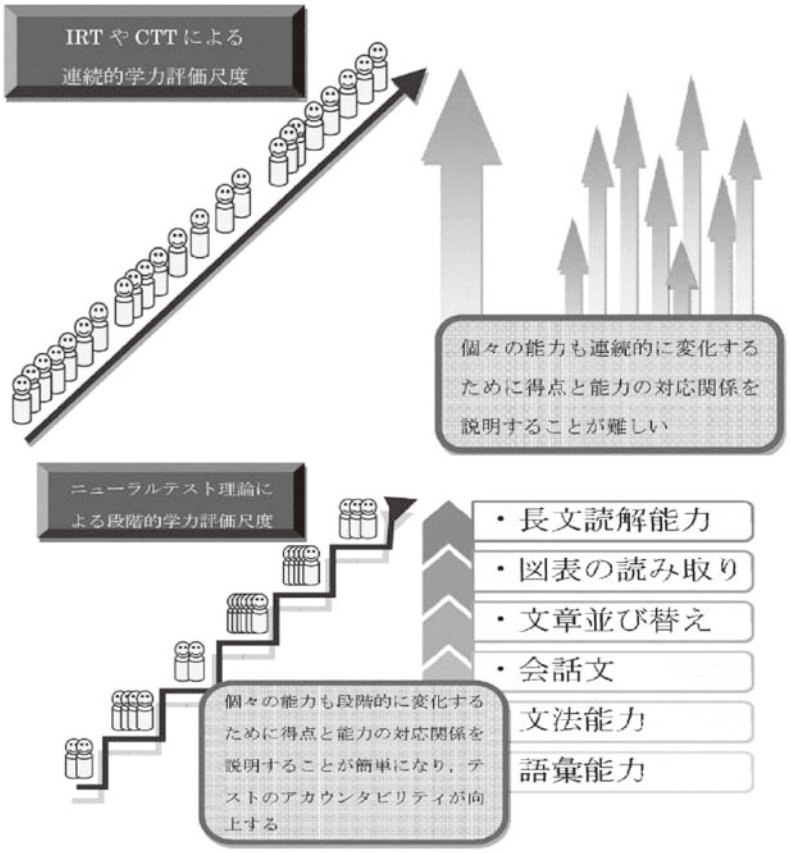

図1　連続的評価尺度と段階的評価尺度

2. 測定精度改善のための情報

正確な測定や適切な評価を行うためには，テストの実施後に，テストを構成する個別の問題（項目）の分析を通して，測定精度の改善に寄与する情報を入手することが可能となる。NTTはテスト結果に関する様々な指標を提供するが，項目参照プロファイル（item reference profile; IRP）はその中の一つに含まれる。IRPはある潜在ランクに所属する受験者がその項目に正答する確率を示し，IRTの項目特性曲線に近いと考えられる。

山川等（2010，p. 79）によれば，IRPでは横軸は受験者の潜在ランクを，縦軸は正答する確率が示されている。図2の場合は，最も下位に属する集団（潜在ランク1）でも，70%以上の確率で正解するような比較的容易な項目であることがわかる。また，ランク5以上が正答する確率に変化があまり見られず，ランク5以上にある受験者の識別力が低い項目であるとも言えよう。図3の項目では，ランク6の周辺で曲線の勾配が変化しており，ランク6以上の集団は正答確率が高くなっており，ランク5以下はあまり差がない。この項目は，ランク6の周辺で識別力が高いと言える。

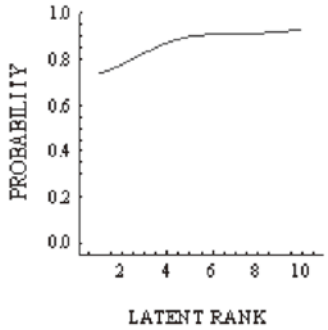

図2　IRP（1）

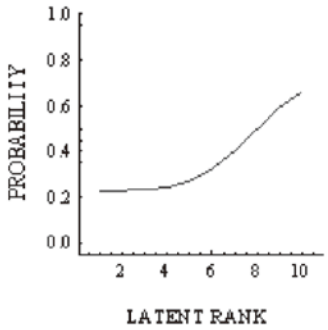

図3　IRP（2）

図4はIRP指標が図示されているIRPのグラフである（木村，2013，p. 20）。項目困難度を表すIRP指標はβとbで，基準となる値（多くの場合0.5としている）に最も近い潜在ランクをβ，その時の値をbとする。項目識別度を表すのがIRP指標αとaである。項目識別度とは上位と下位の受験者の正答確率を分別する程度である。隣り合う2つのIRPの値の差が最大となるペアの小さい方の潜在ランクをα，その時の差がaである。IRP指標γとcは単調増加度を示すもので，隣り合う2つの潜在ランクで正答確率が減少したペア数の割合をγとし，減少した大きさの和がcである。図4中にγ以外のIRP指標を示した（$\beta=5$, $b=0.59$, $\alpha=3$, $a=0.23$, $c=0.10$）。γは，隣り合う2つの潜在ランクで正答確率が減少したのは4ペア中1ペアなので，0.25である。

木村（2013，p. 39）は3つのIRP指標を用いて，テストの測定精度を改善するために，望ましくないテスト項目を抽出する方法を提示している。木村は以下の優先順位を持つ3つの条件に照らし合わせ，1項目ずつ除去し，再分析を行うという指針を提案している。

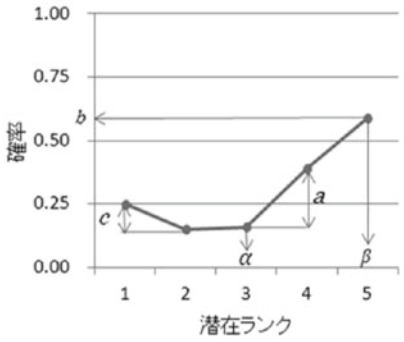

図4　IRPとIRP指標

第1条件：$\gamma = 1$

第2条件：c≧0.05

第3条件：a＜0.05

まず，第1条件に当てはまる項目を探し，複数見つかった場合は第2条件を加え，cの値が最大のものを除去する。第1条件にも第2条件にも当てはまる項目が無くなった場合，第3条件を加え，複数ある場合はa値が最小のものから除去する。

この指針に則って，1項目ずつ望ましくない項目を除去し，再分析を繰り返した結果，除去された13項目の，各分析時点でのIRPとIRP指標が表1に示してある。除去されたセルには網掛けが施されている。

表1　IRPに基づく望ましくない項目の除去

Item	IRP					IRP Index					
	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	α	a	β	b	γ	c
Vgm0047	0.390	0.308	0.243	0.206	0.168	1	0.000	1	0.390	1.000	0.221
Vgm0042	0.266	0.213	0.187	0.154	0.101	1	0.000	1	0.266	1.000	0.165
Vgm0038	0.294	0.241	0.180	0.221	0.261	3	0.041	1	0.294	0.500	0.113
Vgm0008	0.196	0.194	0.186	0.154	0.149	1	0.000	1	0.196	1.000	0.047
Vgm0028	0.253	0.248	0.243	0.215	0.208	1	0.000	1	0.253	1.000	0.045
Vgm0007	0.241	0.275	0.299	0.267	0.205	1	0.034	3	0.299	0.500	0.094
Vgm0061	0.308	0.360	0.318	0.297	0.376	4	0.079	5	0.376	0.500	0.063
Vgm0014	0.919	0.958	0.979	0.958	0.919	1	0.039	1	0.919	0.500	0.060
Vgm0065	0.120	0.125	0.110	0.115	0.115	3	0.005	2	0.125	0.500	0.015
Vgm0032	0.256	0.223	0.221	0.229	0.240	4	0.011	1	0.256	0.500	0.035
Vgm0071	0.151	0.154	0.144	0.163	0.167	3	0.020	5	0.167	0.250	0.011
Vgm0056	0.105	0.133	0.132	0.152	0.165	1	0.028	5	0.165	0.250	0.001
Vgm0053	0.232	0.271	0.282	0.263	0.266	1	0.039	3	0.282	0.250	0.020

木村は，IRPに準拠して除去された項目とIRTに基づいて除去された項目を比較している。IRPの除去指針は記述済みだが，IRTの場合，*MNSQ*＞1.3かつ*Zstd*＞1.96の基準で項目の除去を行っている。*MNSQ*とZ_{STD}とは残差平均と標準化された残差平均を意味する。項目応答理論では，モデルが産出した期待値（各項目に各受験者が正答する確率）と実際のデータの齟齬を残差平均として産出する（吉沢他．2017．p. 55–56）。IRPと同様に，最も望ましくない項目から除去し，1項目を除去するごとに，分析をやり直した結果，8項目が除去された（表2）。

表2　ミスフィット指標による望ましくない項目の除去

	Measure	*SE*	Infit *MNSQ*	Infit *Zstd*	Outfit *MNSQ*	Outfit *Zstd*
Vgm0047	1.28	0.17	1.25	3.06	1.49	3.86
Vgm0042	1.75	0.19	1.18	1.55	1.50	2.81
Vgm0038	1.28	0.16	1.21	2.65	1.48	3.84
Vgm0008	1.65	0.18	1.19	1.74	1.48	2.92
Vgm0028	1.49	0.17	1.18	1.92	1.47	3.19
Vgm0007	1.31	0.16	1.24	2.88	1.49	3.82
Vgm0061	0.92	0.15	1.14	2.31	1.24	2.60
Vgm0014	−2.53	0.25	1.00	0.04	1.49	1.60
Vgm0065	2.39	0.21	1.10	0.69	1.38	1.61
Vgm0032	1.44	0.17	1.16	1.82	1.49	3.45
Vgm0071	2.03	0.19	1.12	1.02	1.30	1.59
Vgm0056	2.28	0.21	1.08	0.62	1.34	1.52
Vgm0053	1.28	0.16	1.18	2.28	1.33	2.74

2つの結果を比較すると，IRT に基づいて除去された8項目は，すべてIRPで除去された項目に含まれている。IRPで除去されたがIRTで除去されていない項目についても，ミスフィット指標[1)]の数値を見ると，どれも除去する基準に近い値のものばかりであった。この結果から，木村はIRPの望ましくない項目の除去方針が妥当であると判断している。

小泉・飯村（2010）は，CTT・ラッシュモデリング・IRPという3つのテスト理論の指標における相関をまとめた（表3）。項目困難度の場合，CTTとラッシュモデリングは−1.00 で完全に一致していた。CTTの困難度とNTTのbの相関は，3ランク（b3）•5ランクの場合（b5）ともに.99と.98で高く，CTTの困難度とNTTのβでは，3ランク（β3）•5ランクの場合（β5）ともに−.61で中程度の負の相関があった。同様にラッシュモデリングの困難度とNTTのb3・b5・β3・β5の相関は−.99，−.98，.61，.61だった。NTTの困難度の指標としてはbとβがあるが，CTT•ラッシュモデリングと似た結果を生み出すという意味ではbの方が適していると言えよう。次に弁別力同士の相関については，CTTの弁別力とNTTの3ランク（a3）•5ランク（a5）での弁別力の

表3　CTT・ラッシュモデリング・NTTの指標の相関

	弁別力	Rasch	α3	a3	β3	b3	γ3	c3	α5	a5	β5	b5	γ5	c5
困難度	−.37	−1.00	−.43	−.54	−.61	.99	.27	.06	−.54	−.56	−.61	.98	.08	.23
弁別力	—	.37	.13	.90	.14	−.42	−.50	.29	.05	.90	.08	−.41	−.45	.23
Rasch b		—	.43	.55	.61	−.99	−.27	−.06	.54	.57	.61	−.98	−.08	−.23
α3			—	.24	.75	−.40	.10	.18	.70	.27	.74	−.35	.22	−.05
a3				—	.28	−.60	−.49	.27	.21	.97	.24	−.57	−.37	.12
β3					—	−.58	.16	.08	.71	.29	.96	−.55	.27	−.10
b3						—	.30	.02	−.53	−.61	−.60	.98	.10	.20
γ3							—	−.62	.26	−.47	.17	.28	.75	−.31
c3								—	−.01	.24	.12	.05	−.45	.59
α5									—	.23	.73	−.50	.33	−.15
a5										—	.25	−.59	−.31	.04
β5											—	−.56	.30	−.11
b5												—	.09	.22
γ5													—	−.75

注．困難度＝CTTの困難度。Rasch b＝ラッシュモデリングでの困難度。α3＝3ランク時のα。.22〜.27までの相関はp＜.05で，.28以上の相関はp＜.01で有意（両側検定）だった。

相関はともに.90だった。

以上のCTTとNTT（b）の項目困難度間（表4）と，CTTとNTT（a）の弁別力間の相関（表5）をまとめている。

それに基づくと，Brown（2005）の基準で適切な項目困難度は.30から.70の間だが，CTTの.30

表4　CTTとNTT（b）との対応

		3ランク				5ランク			
	k	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
.90〜1.00	28	.91	.07	.78	1.00	.90	.08	.76	1.00
.80〜.90	14	.72	.05	.63	.80	.69	.06	.58	.78
.70〜.80	9	.61	.07	.50	.75	.58	.09	.43	.74
.60〜.70	4	.49	.08	.37	.55	.51	.03	.48	.53
.50〜.60	9	.50	.07	.40	.61	.50	.04	.43	.56
.40〜.50	6	.42	.03	.38	.48	.49	.04	.45	.57
.30〜.40	9	.30	.05	.25	.37	.41	.03	.36	.45
.20〜.30	10	.22	.04	.17	.30	.29	.05	.21	.36
.10〜.20	1	.16	—	—	—	.16	—	—	—

注．*k*＝項目数。*Min*＝最小値。*Max*＝最大値。

表5　CTTとNTT（a）との対応

		3ランク				5ランク			
	k	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
.50〜.60	7	.31	.05	.26	.39	.21	.04	.16	.26
.40〜.50	21	.22	.05	.12	.32	.16	.04	.10	.24
.30〜.40	25	.16	.05	.08	.27	.12	.04	.06	.22
.25〜.30	9	.14	.04	.07	.18	.09	.03	.04	.12
.20〜.25	8	.07	.03	.04	.13	.05	.02	.02	.09
.10〜.20	12	.05	.03	.01	.09	.04	.02	.01	.07
.00〜.10	8	.01	.02	.00	.04	.01	.01	.00	.03

注．*k*＝項目数。*Min*＝最小値。*Max*＝最大値。

はNTTでは.30〜.40程度，CTTの.70はNTTで.55〜.65程度と言える判断している。弁別力の関係については，Henning（1987）の.25以上という基準で考えると，CTTの.25はNTTで.05から.15程度という数字が基準の候補として挙げられると指摘している。

3. 受験者評価のための情報

3.1　テスト参照プロファイル（test reference profile; TRP）

TRPは，項目参照プロファイルに重みをつけた和であり，各潜在ランクに属する受験者の期待得点を表している（山川等，2010，p. 79）。図5のX軸は項目参照プロファイルと同じく潜在ランクを示し，Y軸は期待得点を示している。ランク1に属する受験者は，11問程度に正解するし，ランク10の者は23問程度に正解すると解釈できる。潜在ランクが上がるごとに確率が上昇することを「単調増加」と言う。TRPで単調増加することを「弱順序配置条件」，IRPで単調増加することを「強順序配置条件」と呼び，NTTでは弱順序配置条件が前提になっている（Shojima,

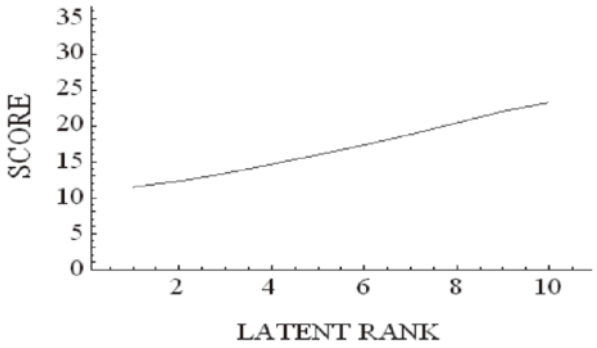

図5　TRP

2008）。

3.2　ランク・メンバーシップ・プロファイル（rank membership profile; RMP）

RMPとは各受験者が，それぞれの潜在ランクに所属する確率である。木村（2013，p. 20）によれば，潜在ランクの推定値が同じでも，RMPが異なると受験者に別の視点からフィードバックを与えられる。例えば，図6と図7は同じテストから得られた2人の受験者のRMPである。この2人の潜在ランクは同じ3だが，図7の受験者は潜在ランク4への所属確率も高いので，潜在ランク3から4に移行しつつあると考えられる。

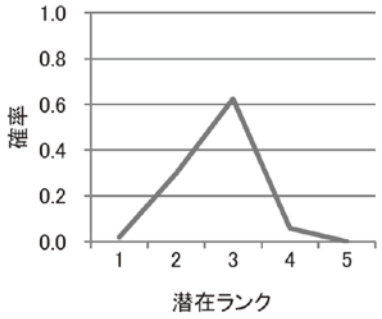

図6　RMP（1）

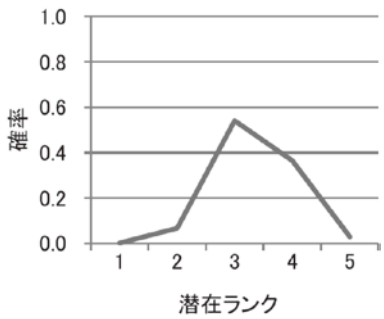

図7　RMP（2）

図6の受験者は，「現在のところ5段階中のランク3だが，ランク2にも近い状況である。易しい問題の中にもできないところがあると思われるので，より難しい問題に取り組む前に，易しい問題の中で不得意なところを復習するとよいだろう」というフィードバックが考えられる。一方，同じランク3に所属する図7の受験者には，「現在のところ5段階中のランク3だが，一つ上のランク4にも近い状況である。基礎的な問題はほぼマスターしていると思われるので，より難しい問題に積極的にチャレンジするとよいだろう」というフィードバックが考えられる。

また，木村（2013，p. 21）によると，同一の学習者のRMPを時系列で示して学力の変化を見

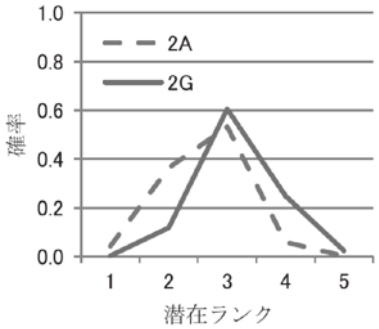

図8　同一学習者のRMPの変化

せることができる。図8はテスト2Aから3か月後に実施したテスト2GのRMPを表す。この図を受験者に見せて，「3か月前と所属するランクは3で変わりないが，前回に比べて易しい問題に正解できるようになっている。今後は，少し難しい問題にもチャレンジしていくとよいだろう」といったフィードバックを与えられる。

木村（2013, p. 93）は，テストが提供するこのような診断的情報に関して次のように述べている。

> 何らかの学習・教育の後に行われ，学習の成果を評価するテストの場合，どのようなことができるようになったのか，まだできるようになっていないことは何か，といった情報を提供することが望まれる。つまり，テストを実施した段階での，そのテストが測定しようとしている潜在能力について診断的情報を提供できることが望ましい。テストとそれに基づく学習評価の関係は，前者が診断的情報を提供することで後者の説明内容に具体性が生まれるようになることが健全であり，そのような関係が築けるように教育者とテスト開発者は努力すべきである．

3.3　ランク分けのための受験者情報

大学入学後にクラスを能力別に分ける時には，一般的にプレイスメントテストの結果に基づいてクラス分けが行われている。その際にテスト得点を基準に学生を分割してその所属クラスを決定するのが通例行われている方法であろう。

木村（2013，p. 62）は英語プレイスメントテストの結果をNTT・IRT・CTTという3つのテスト理論を用いて能力別クラス分けを行った。各クラスの英語学力を産出するのに，NTTでは潜在ランク（R_T），IRTでは推定能力（θ_T），CTTでは正答数（S_T）が用いられた。各能力別クラスの英語力をまとめたのが表6である。

表6　各クラスの英語学力（R_T，θ_T，S_T）の代表値と散布度の比較

Class	*n*	R_T		θ_T		S_T	
		Mdn	*Range*	*M*	*SD*	*M*	*SD*
Class 01	15	6	5	−3.06	0.604	26.9	3.88
Class 02	15	11	5	−1.39	0.584	35.5	3.76
Class 03	16	17	4	0.11	0.652	42.6	3.66
Class 04	14	21	3	0.97	0.698	46.9	3.09
Class 05	15	26	6	2.89	1.204	54.3	4.84

R_T，θ_T，S_Tの3つともクラス間に有意差（$p<.01$）があった。多重比較を行ったところ，R_Tについては連続するクラス間以外で，θ_TとS_Tについてはすべてのクラス間で有意差（$p<.01$）が認められた。さらに，Rとθの間の相関係数は，.96という高い値を示した。SとRの相関係数も.94，Sとθの相関係数も.96の高い値を示した（表7参照）。

表7　R_T，θ_T，S_Tの相関

総合評価	R_T	θ_T	S_T
R_T	—	.96	.94
θ_T		—	.96
S_T			—

これらの状況から，NTTの潜在ランク，IRTの潜在能力値，CTTの正答数，いずれで評価しても，能力はほぼ同じ順位で評価されることがわかった。ただし，IRTの連続した間隔尺度上で表現される細かい数値は，クラス分けのためには不要であるばかりか，どこに閾値を設けるかの判断が難しい。それに対し，NTTの潜在ランクは，はじめから順序尺度上に限られた数の段階に分けて表現されるので，クラスをどこで分けるかの判断を容易に行うことができる。したがって，プレイスメントテストの分析にIRTを活用することは，妥当かつ有用であると言える。

3.4　ランク分けの「テスト適合度」

上記の3.3で示したように，NTTはクラス分けに有用な情報を提供してくれることがわかった。そもそも，潜在ランクは最尤法[2)]やベイズ法[3)]で決められるが（Shojima, 2007a, 2007b），受験者の学力を分ける潜在ランク数はいくつでも自由に設定することができる。では，データに最も合致した潜在ランク数を決定するための指標はないのであろうか。

小泉・飯村（2010）によると，NTTでは設定したランク数のモデルとデータとの一致度を示す「テスト適合度」が2種類算出される。分析したサンプルについてだけ記述する時には「テスト適合度」

表8　ランク・メンバーシップ・プロファイル（RMP）に基づくテスト適合度

	X^2	df	p	NFI	RFI	CFI	RMSEA	AIC	CAIC	BIC
基準		.05以上				.90以上	.05未満	低い値がよりよい		
NTT2	962.26	4320	1.00	0.54	0.54	1.00	0.00	−7677.74	−24916.41	−20596.41
NTT3	637.82	4230	1.00	0.69	0.69	1.00	0.00	−7822.18	−24701.71	−20471.71
NTT4	539.70	4140	1.00	0.73	0.73	1.00	0.00	−7740.30	−24260.69	−20120.69
NTT5	470.66	4050	1.00	0.76	0.76	1.00	0.00	−7629.34	−23790.59	−19740.59
NTT6	425.17	3960	1.00	0.78	0.78	1.00	0.00	−7494.83	−23296.95	−19336.95

注．NTT2＝NTTで2グループに分けた場合。NFI＝normed fit index; RFI＝relative fit index; CFI＝comparative fit index; RMSEA＝root mean square error of approximation; AIC＝Akaike information criterion; CAIC＝consistent AIC; BIC＝Bayes information criterion. AIC，CAIC，BICは負の値になりえ，（絶対値ではなく）そのままの値で小さい方がよいモデルと判断する（例：−50，−100なら後者が小さく，よりよいモデルを示す）。適合度指標の詳細については豊田（2007）参照。IFI（incremental fit index）・TLI（Tucker-Lewis index）も算出されたが，CFIと同じ値だった。

が使用されるのに対して，母集団を仮定し，母集団について言及したい時にはもう一つの「RMPに基づくテスト適合度」（RMP＝rank membership profile）が用いられる。テスト適合度を判断する際には，複数の適合度指標の結果を参照する。表8は，RMPに基づくテスト適合度での適合度指標の結果である。小泉・飯村はランクを2から6まで分析した。CFI[4] とRMSEA[5] を見ると2から6ランクまで全般的に適合度は高かったが，AIC[6]・CAIC[7]・BIC[8] の値を見ると，すべて3グループの時が最も低く，テスト適合度が高かった。そのため，NTT の場合は受験者を3群まで分けることができることが示された。

3.5　Can-do-Statementsとの相性

荘島（2010，p. 106）は，テストは単に能力を数値化して，受験者に点数を与えるだけでは説明責任を果たしていないと考える。テストの結果を通して，受験者の具体的な能力を説明でき，それらの細目の獲得の程度を知らせることができなければ，試験としての要件を満たさないとする。テストは，受験者がどの程度の能力を有しているかについての診断的機能を有していることが望ましい。すなわち，学力を段階評価して各能力段階の細目を明らかにすることが重要である。そのような能力カタログをCan- Do Statements（CDS）と言う。CDSは連続尺度のもとで作成するのは困難で，順序尺度の方が各段階に対応する能力水準を記述しやすい。

3.6　学力層ごとの誤答分析

松宮（2012）は誤答分析を行うに際しては，受験者の誤答類型を考察することは指導に活用できる情報として有用であると考える。とりわけ，学力層による誤答選択状況まで踏み込んで分析すれば，より詳細な情報が得られる。一般的に学力に対する正答選択率は単調増加である。すなわち，学力の高い受験者ほど正答率が高くなるはずである。しかしながら，だからといってすべての誤答選択肢の選択率が単調減少するわけではない。ある学力層に偏って好まれる誤答選択肢がある。このような層を特定することができれば，指導すべき層を絞ることができ，具体的で有

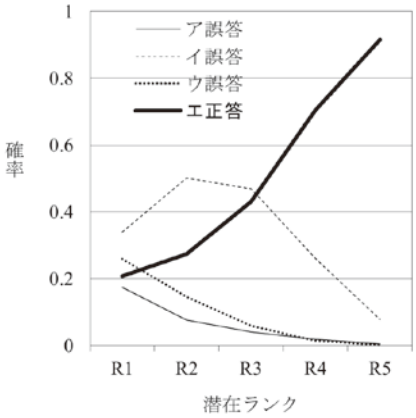

図9　ICRP（項目3）

効な手立ても考えやすい。

NTTでは項目カテゴリ参照プロファイル（Item Categories Reference Profile ;ICRP）がそのような情報を提供できる。ICRPは選択肢ごとの潜在ランク別選択確率を表現されるので，どの学力層がどの誤答を選択する確率が高いかを検討できる。例えば，図9の場合，この項目3の問題では選択肢アとウの誤答率は潜在ランク（学力）が高いほど下がっているので問題はない。しかし，選択肢イに関して，R2（ランク2）・R3（ランク3）の学力層で誤答イの選択率がR1（ランク1）よりも増えている。つまり，ランク2・3と学力が上がるにつれて本来下がるべき誤答率が逆に上がっている。この学力層の受験者に対して，誤答の要因を突き止めて，適切な指導が必要であることが明らかになっている。

このように，ICRPを検討することで，全体から見た誤答率だけでなく，学力層に注目した誤答情報を用いて指導が見直されるならば，内容によって指導のターゲットを絞るという点で，一人一人の学習者にカスタマイズされたより的確な改善が期待できる。

4. まとめにかえて

本稿では，ニューラルテスト理論（NTT）を用いたテスト結果のフィードバックに関する先行研究を概観した。NTTは測定と評価にかかわる様々なシーンで有用な情報を提供できることが明らかとなった。今後の課題として，同理論の枠組みを用いたテスト結果の分析を試みてみたいと考えている。その結果と古典的テスト理論を用いた分析結果の比較を通して，NTTの妥当性の検証を進めてみたい。

注

1） ミスフィット指標にはInfitとOutfitがある。InfitとはInformation Weighted mean square fit statisticsの略で個々の応答の分散をもとにした加重平均である。他方，Outfitとはoutlier sensitive mean square fit statisticsの略で，標準化残差の和の単純平均である（靜，2007，p. 312–313）。

2） 未知のパラメーターを推定する際に，実際に与えられたデータは，パラメーターがどういう値のときに最も生じやすいかを調べて，そのデータが生じる確率が最大になるようなパラメーター値をもって，推定値とする方法（大友，1996，p. 122）

3） 未知である母数の事前情報を考慮して，その推定値を求める方法（豊田，2007，p. 150）

4） Comparative Fit Indexの略。分析しているモデルが独立モデルから飽和モデルまでのどのあたりに位置するかを示す。1に近いほど望ましく，0.90より大きいとよいモデルと言われる（小塩，2008，p. 110–111）。

5） Root Mean Square Error of Approximationの略。RMSEは小さいほど望ましい。一般に，0.05以下であればよく，0.10以上はよくないとされる（小塩，2008，p. 110–111）。

6） Akaike’s Information Criterionの略。絶対的な基準があるわけではないが，小さな値をとるものほどよいと判断するための指標（小塩，2008，p. 110–111）。

7） Consistent Akaike Information Criterion。AICのサンプルサイズの補正をさらに加えたもの。

8） Bayesian Information Criterionの略。小さな値をとるものほどよい。

参考文献

Brown, J. D.（2005）. Testing in language program: A comprehensive guide to English language assessment. New York: McGraw-Hill.

Henning, G.（1987）. A guide to language testing: Development, evaluation, research. Boston, MA: Heinle & Heinle

木村哲夫（2013）.『潜在ランク理論を用いたコンピュータ適応型テストのためのアルゴリズムの提案と実装』早稲田大学審査学位論文.

木村哲夫（2016）.「潜在ランク理論」『日本言語テスト学会20周年記念特別号』217–222.

小泉利恵・飯村英樹（2010）.「ニューラルテスト理論の特徴：古典的テスト理論・ラッシュモデリングとの比較から」『JLTA（Japan Language Testing Association）Journal』13, 91–109.

小塩真司（2008）. はじめての共分散構造分析―Amosによるパス解析, 東京図書

松宮功（2012）. 項目カテゴリ参照プロファイルによる学力テストの誤答分析. 京都府総合教育センター研究紀要第2集, p. 32–36

中村洋一（著）・大友賢二（監修）（2002）.『テストで言語能力は測れるか―言語テストデータ分析入門』東京：桐原書店

大友賢二（1996）.『項目応答理論入門』東京：大修館書店

清水裕子・木村真治・杉野直樹　山川健一・大場浩正・中野美知子（2003）. 英語文法能標準テストの妥当性・信頼性の検証と 新英語文法能力テスト Measure of English Grammar（MEG）政策科学10–3, 59–68

靜哲人(2007).『基礎から深く理解するラッシュモデリング―項目応答理論とは似て非なる測定のパラダイム』, 関西大学出版部

荘島宏二郎（2010）.「ニューラルテスト理論―学力を段階評価するための潜在ランク理論―」植野真臣・荘島宏二郎（編著）『学習評価の新潮流』東京: 朝倉書店.

Shojima, K.（2007a）. Bayesian estimation of latent rank in neural test theory. DNC Research Note, 07–15. Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/jpaper.htm

Shojima, K.（2007b）Maximum likelihood estimation of latent rank under the neural test model. DNC Research Note, 07–04. Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/jpaper.htm

Shojima, K.（2008）. Neural test theory: A latent rank theory for analyzing test data. DNC Research Note, 08–01. Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/jpaper.htm

豊田秀樹（2007）.『共分散構造分析―Amos編』東京：東京書籍

山川健一・杉野直樹・清水裕子・大場浩正・中野美知子（2010）.「ニューラルテスト理論の第二言語習得研究への応用可能性1」www.rd.dnc.ac.jp/~shojima/ntt/YamakawaAAAL2010.pdf

吉澤清美・髙瀬敦子・大槻きょう子（2017）. 日英バイリンガル文法テストの開発：多読とのかかわりにおける文法能力の発達を測定する 関西大学外国語学部紀要. 第16号. 45–60